Published online 18 November 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D71–D75***
doi:10.1093/nar/gkr981

Reorganizing the protein space at the Universal Protein Resource (UniProt)

The UniProt Consortium[1,2,3,4,*]

[1]The EMBL Outstation, The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [2]SIB Swiss Institute of Bioinformatics, Centre Medical Universitaire, 1 rue Michel Servet, 1211 Geneva 4, Switzerland, [3]Protein Information Resource, Georgetown University Medical Center, 3300 Whitehaven St NW, Suite 1200, Washington, DC 20007 and [4]University of Delaware, 15 Innovation Way, Suite 205, Newark, DE 19711, USA

Received September 29, 2011; Accepted October 14, 2011

ABSTRACT

The mission of UniProt is to support biological research by providing a freely accessible, stable, comprehensive, fully classified, richly and accurately annotated protein sequence knowledgebase, with extensive cross-references and querying interfaces. UniProt is comprised of four major components, each optimized for different uses: the UniProt Archive, the UniProt Knowledgebase, the UniProt Reference Clusters and the UniProt Metagenomic and Environmental Sequence Database. A key development at UniProt is the provision of complete, reference and representative proteomes. UniProt is updated and distributed every 4 weeks and can be accessed online for searches or download at http://www.uniprot.org.

INTRODUCTION

The specific aim of UniProt is to provide a centralized repository of protein sequences with comprehensive coverage and a systematic approach to protein annotation, incorporating, interpreting, integrating and standardizing data from a large number of disparate sources. It is the most comprehensive catalog of protein sequence and functional annotation and has four components optimized for different uses. The UniProt Knowledgebase (UniProtKB) is an expertly curated database, a central access point for integrated protein information with cross-references to multiple sources. The UniProt Archive (UniParc) (1) is a comprehensive sequence repository, reflecting the history of all protein sequences. UniProt Reference Clusters (UniRef) (2) merge closely related sequences based on sequence identity to speed up searches while the UniProt Metagenomic and Environmental Sequences database (UniMES) was created to respond to the expanding area of metagenomic data. With the exponential increase in sequence data, it becomes critically important to structure the data in an automatic fashion that will provide a global genome/proteome and gene-product-centric display of the sequence space that allows drilling down to variations and annotations for each specific sequence. UniProt is reorganizing the UniProtKB data representation to achieve this and to facilitate an optimal use of the wealth of sequence and functional data available at the protein level.

NEW AND ONGOING DEVELOPMENTS

Organizing the protein sequence space

With the significant increase in the number of complete genomes sequenced, it is essential to organize these data in a way that allows users to effectively navigate the growing number of available complete proteome sequences. The approach adopted by UniProt to meet this challenge is to create a UniProtKB core subset according to an evolving set of criteria. This will ensure that users will only find the most relevant and best annotated sequences when searching instead of drowning in reports of redundant sequences. Those redundant sequences will continue to be provided in the UniProtKB non-core subset.

The UniProtKB core consists of reference and representative proteomes (RPs) which are a subset of the complete proteome sets and the manually reviewed UniProtKB/Swiss-Prot section, thus providing both completeness and expert literature curation while eliminating redundancy.

Complete proteomes. A complete proteome is defined as the entire set of proteins expressed by a specific organism. The majority of the UniProt complete proteome sets are

*To whom correspondence should be addressed. Tel: +44 1223 494435; Fax: +44 1223 494468; Email: apweiler@ebi.ac.uk

based on the translation of a completely sequenced genome, and will normally include sequences that derive from extra-chromosomal elements such as plasmids or organellar genomes in organisms where these occur. Some complete proteomes may also include protein sequences based on high quality cDNAs that cannot be mapped to the current genome assembly (due to sequencing errors or gaps). These are only included in the complete proteome following manual review of the supporting evidence, including careful analysis of homologous sequences from closely related organisms.

UniProt complete proteome sets may include both manually reviewed (UniProtKB/Swiss-Prot) and unreviewed (UniProtKB/TrEMBL) entries. The proportion of reviewed entries varies between proteomes, and is obviously greater for the proteomes of intensively curated model organisms: some complete proteomes, such as those of *Saccharomyces cerevisiae 288C* and *Escherichia coli strain K12*, consist entirely of reviewed entries. For *Homo sapiens*, the UniProt set provides reviewed entries for each known gene and unreviewed entries for some isoforms. Curation is a continuing process, and complete proteome sets are updated in a regular manner as new information becomes available: pseudogenes and other dubious uncharacterized ORFs may be removed while other newly identified and characterized sequences may be added.

Currently, the majority of UniProt complete proteomes are based on translations of genome sequence submissions to the International Nucleotide Sequence Database Consortium (INSDC) (3). A complementary pipeline for import of protein sequences has been developed in collaboration with Ensembl (4) that provides proteome sequences for a number of key genomes of special interest that currently lack a complete INSDC submission. As this pipeline covers organisms for which we already have some sequences in UniProtKB, these existing sequences have to be reconciled with those imported. The procedure works in the following way:

(1) Ensembl sequences are first mapped to their UniProtKB counterparts under stringent conditions, requiring 100% identity over 100% of the length of the two sequences;
(2) Ensembl sequences that are absent from UniProtKB are imported into UniProtKB/TrEMBL and tagged with the keyword 'Complete proteome'; and
(3) A complete proteome is formed from all UniProtKB/Swiss-Prot entries (irrespective of whether they map to Ensembl) plus only those UniProtKB/TrEMBL entries mapping to Ensembl.

These complete proteomes can then be additionally defined as reference and RPs and as such are included in the UniProtKB core.

Reference proteomes. UniProt has defined a set of reference proteomes which are 'landmarks' in proteome space. Reference proteomes have been selected to provide broad coverage of the tree of life, and constitute a representative cross-section of the taxonomic diversity to be found within UniProtKB. They include the proteomes of well-studied model organisms [including those in the now defunct IPI sets (5)] and other proteomes of interest for biomedical and biotechnological research. These are the proteomes which are preferentially selected for manual curation when resources permit. Species of particular importance may be represented by numerous reference proteomes for specific ecotypes or strains of interest.

Currently, UniProt has defined 455 reference proteomes in close collaboration with Ensembl and NCBI Reference Sequence collection (RefSeq) (6). The collaboration's goal is that the same consensus sets are provided by all three resources. The reference proteome set will be continuously reviewed as new proteomes of interest become available and as existing taxonomic classifications are revised. We would very much welcome interaction with our user community on our current list of reference proteomes and suggestions for new candidates.

Representative proteomes. There are hundreds of complete proteomes not included in the UniProt Reference Proteomes and this number is expected to increase many fold with sequences from new organisms as well as additional isolates and strains of existing organisms. This flood of new proteomes will decrease the sensitivity of sequence and text searches. To help cope with this, we are working on a computationally derived set of RPs. A RP is the proteome that can best represent all the proteomes in its group in terms of the majority of the sequence space and annotation (7). Each RP is selected from a RP group (RPG) containing similar proteomes calculated based on co-membership in UniRef50 clusters. The most information-rich proteome (based on a collective annotation score of its entries) is selected as the 'Representative' from each RPG. If a reference proteome exists in a cluster, it will be selected. RPs are calculated at 75, 55, 35 and 15% co-membership thresholds using a top–down approach that ensures an RP at a lower threshold is also an RP at a higher threshold. The 55% threshold (RP55) will be used in the 'core' set as it most closely follows standard taxonomic classifications, and preserves the majority of the annotation and sequence diversity of the entire UniProtKB, while reducing the sequence space by more than 80%.

Access and availability. The complete and reference proteomes were available on the UniProt web and ftp sites from September 2011 and it is planned that the UniProtKB core set will be available by the end of the year for FTP download, similarity searches and searching or browsing on the web site in our new complete proteomes portal which is currently under development at http://www.uniprot.org/taxonomy/completeproteomes. New keywords 'Reference proteome' and 'RP' have been created to allow the easy retrieval of proteome sets. This new portal will provide users with information and simple statistics for both complete proteomes and their individual components, such as chromosomes and plasmids.

New biocuration pages on UniProt website

An integral part of the UniProtKB core is the expertly manually curated subset. The biocuration process involves the integration and interpretation of information from a variety of sources as well as accurate and comprehensive representation of the data. It adds a wealth of information to UniProtKB records including information related to the role of a protein such as its function, structure, subcellular location, interactions with other proteins and domain composition, as well as a wide range of sequence features such as active sites and post-translational modifications. Manual curation provides high-quality data for experimentally characterized proteins and consists of a critical review of experimental and predicted data for each protein as well as manual verification of each protein sequence. This information is included in the manually reviewed Swiss-Prot section of UniProtKB. In response to the ever-increasing amounts of sequence data, automated methods have been developed by the UniProt Consortium to annotate uncharacterized proteins with a high degree of accuracy and these methods are used to enhance the unreviewed records in UniProtKB/TrEMBL by enriching them with automatic classification and annotation. In order to keep UniProt users informed of curation practices and priorities within the project, the UniProt website has been updated to include a new section describing UniProt biocuration at http:/www.uniprot.org/help/biocuration. This section provides an overview of the manual curation process including a standard operating procedure (SOP) as well as details of current manual curation priorities. In addition, information is provided about the automatic annotation systems developed and used within the group. Additional useful information such as statistics, links to related resources and relevant publications are also provided. The pages will continue to be updated on a regular basis to provide users with the latest information about the UniProt curation process and activities.

UniProt outreach

In order to better meet users' demands and wishes, we held two UniProt interactive workshops at PIR and EBI in June and September 2011. This is part of a larger effort by the UniProt Consortium to focus on its user experience. UniProt users from academia and industry, mostly from a wet lab environment, were invited to help us evaluate UniProt's databases and web site usability over a 2-day period. The activities were designed to understand why and how people use UniProt, gaps and usability issues and to identify users' priorities and requirements to help guide future development. The workshops proved to be a success and the feedback material will be used to improve data visualization, web site usability and data provision in the near future. We will also be increasing our training and dissemination activities and testing prototypes that include new designs intended to address the findings of the workshops.

DATABASE ACCESS AND FEEDBACK

The www.uniprot.org website (8) is the primary access point to our data and documentation and to tools such as full text and field-based text search, sequence similarity search, multiple sequence alignment, batch retrieval and database identifier mapping. These tools can be accessed directly through a tool bar that appears at the top of every page. Most data (including documentation and help) can be searched through the full text search, which allows searches requiring no prior knowledge of our data or search syntax. Results are sorted by relevance and, where possible, suggestions are provided to help refine searches that yield too many or no results. The field-based text search supports more complex queries. These can be built iteratively with the tool bar's query builder or entered manually in the query field, which can be faster and more powerful (www.uniprot.org/help/text-search). Searching with ontology terms is assisted by auto-completion, and we also provide the possibility of using ontologies to browse search results. Viewing of result sets, as well as database entries, is configurable. Sequence similarity search results which have been adapted to the new EMBL-EBI framework for web services tools (9) can be filtered by taxonomy to gain a quick overview of the taxonomic distribution of the results. The sequence annotations of matched UniProtKB entries can be projected onto the sequence alignments to see at a glance if important positions are conserved. Columns can be added to or removed from the result table to see more functional annotation than is available in the default view. The site has a simple and consistent URL scheme and all searches can be bookmarked to be repeated at a later time. The home page features a site-tour as a quick introduction for novice users. In response to user requests for various downloadable data sets (e.g. all reviewed human entries in FASTA format), we have removed all download limits to allow this functionality by directly querying the website. However, large downloads are given low priority in order to ensure that they do not interfere with interactive queries, and they can therefore be slow compared to downloads from the UniProt FTP server. We therefore recommend downloading complete datasets from ftp .uniprot.org/pub/databases. The website offers various download formats which depend on the chosen dataset (e.g. plain text, XML, RDF, FASTA, GFF for UniProtKB). The columns of result tables can be configured for customized downloads in tab-delimited or Excel format. All data is available in RDF (www.w3.org/RDF/), a W3C standard for publishing data on the Semantic Web. Programmatic access to data and search results is possible via simple HTTP (REST) requests (www.uniprot.org/faq/28). Java applications can also make use of our Java API (UniProtJAPI) (10).

While the UniProt website provides a query interface for all UniProt data, users frequently require the facility to search across related data in different databases. BioMart is an open source query-oriented data management system that allows for integrated querying of biological data resources regardless of their geographical

locations. A UniProt Biomart (http://www.ebi.ac.uk/uniprot/biomart/martview) is available which allows complex queries between UniProt and other data resources such as PRIDE (11), Ensembl and InterPro (12).

We are constantly trying to improve our databases and services in terms of accuracy and representation and hence, consider your feedback extremely valuable. Please contact us if you have any questions via www.uniprot.org/contact or email us directly at help@uniprot.org. Information is provided at www.uniprot.org/help/submissions about data submissions and updates. Extensive documentation on how to best use our resource is available at www.uniprot.org/help/. UniProt is freely available for both commercial and non-commercial use. Please see www.uniprot.org/help/license for details. New releases are published every 4 weeks except for UniMES, which is updated only when the underlying source data are updated. Statistics are available with each release at www.uniprot.org.

ACKNOWLEDGEMENTS

UniProt has been prepared by: Rolf Apweiler, Maria Jesus Martin, Claire O'Donovan, Michele Magrane, Yasmin Alam-Faruque, Ricardo Antunes, Elisabet Barrera Casanova, Benoit Bely, Mark Bingley, Lawrence Bower, Borisas Bursteinas, Wei Mun Chan, Gayatri Chavali, Alan Da Silva, Emily Dimmer, Ruth Eberhardt, Francesco Fazzini, Alexander Fedotov, John Garavelli, Leyla Garcia Castro, Michael Gardner, Reija Hieta, Rachael Huntley, Julius Jacobsen, Duncan Legge, Wudong Liu, Jie Luo, Sandra Orchard, Samuel Patient, Klemens Pichler, Diego Poggioli, Nikolas Pontikos, Sangya Pundir, Steven Rosanoff, Tony Sawford, Harminder Sehra, Edward Turner, Tony Wardell, Xavier Watkins, Matt Corbett, Mike Donnelly, Pieter van Rensburg, Mickael Goujon, Hamish McWilliam, and Rodrigo Lopez at the European Bioinformatics Institute (EBI); Ioannis Xenarios, Lydie Bougueleret, Alan Bridge, Sylvain Poux, Nicole Redaschi, Ghislaine Argoud-Puy, Andrea Auchincloss, Kristian Axelsen, Delphine Baratin, Marie-Claude Blatter, Brigitte Boeckmann, Jerven Bolleman, Laurent Bollondi, Emmanuel Boutet, Silvia Braconi Quintaje, Lionel Breuza, Edouard deCastro, Lorenzo Cerutti, Elisabeth Coudert, Beatrice Cuche, Isabelle Cusin, Mikael Doche, Dolnide Dornevil, Severine Duvaud, Anne Estreicher, Livia Famiglietti, Marc Feuermann, Sebastien Gehant, Serenella Ferro, Elisabeth Gasteiger, Vivienne Gerritsen, Arnaud Gos, Nadine Gruaz-Gumowski, Ursula Hinz, Chantal Hulo, Nicolas Hulo, Janet James, Silvia Jimenez, Florence Jungo, Thomas Kappler, Guillaume Keller, Vicente Lara, Philippe Lemercier, Damien Lieberherr, Xavier Martin, Patrick Masson, Madelaine Moinat, Anne Morgat, Salvo Paesano, Ivo Pedruzzi, Sandrine Pilbout, Monica Pozzato, Manuela Pruess, Catherine Rivoire, Bernd Roechert, Michel Schneider, Christian Sigrist, Karin Sonesson, Sylvie Staehli, Eleanor Stanley, Andre Stutz, Shyamala Sundaram, Michael Tognolli, Laure Verbregue, and Anne-Lise Veuthey at the SIB Swiss Institute of Bioinformatics (SIB); Cathy H. Wu, Cecilia N. Arighi, Leslie Arminski, Winona C. Barker, Chuming Chen, Yongxing Chen, Pratibha Dubey, Hongzhan Huang, Abhishek Kukreja, Kati Laiho, Raja Mazumder, Peter McGarvey, Darren A. Natale, Thanemozhi G. Natarajan, Natalia V. Roberts, Baris E. Suzek, C. R. Vinayaka, Qinghua Wang, Yuqi Wang, Lai-Su Yeh and Jian Zhang at the Protein Information Resource (PIR).

FUNDING

National Institutes of Health (NIH) grant (1U41HG006104-02; the EBI's involvement in UniProt comes from the European Commission contract SLING grant (226073); the National Institutes of Health grant (2P41HG02273-07); the British Heart Foundation (SP/07/007/23671); Kidney Research UK (KRUK) (RP26/2008); UniProtKB/Swiss-Prot activities at the SIB are supported in addition from the Swiss Federal Government through the Federal Office of Education and Science and from the European Commission contracts GEN2PHEN (200754), MICROME (222886-2) and SLING (226073); PIR activities are also supported by National Institutes of Health (NIH) grants 5R01GM080646-05, 3R01GM080646-04S2, 2R01GM080646-06, 5G08LM010720-02, 3P20RR016472-09S2 and National Science Foundation (NSF) grants DBI-0850319 and DBI-1062520. Funding for open access charge: National Institutes of Health (NIH) grant (1U41HG006104-02).

Conflict of interest statement. None declared.

REFERENCES

1. Leinonen,R., Diez,F.G., Binns,D., Fleischmann,W., Lopez,R. and Apweiler,R. (2009) UniProt archive. *Bioinformatics*, **20**, 3236–3237.
2. Suzek,B.E., Huang,H., McGarvey,P., Mazumder,R. and Wu,C.H. (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, **23**, 1282–1288.
3. Brunak,S., Danchin,A., Hattori,M., Nakamura,H., Shinozaki,K., Matise,T. and Preuss,D. (2002) Nucleotide sequence database policies. *Science*, **298**, 1331–1332.
4. Hubbard,T.J.P., Aken,B.L., Ayling,S., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Clarke,L. *et al.* (2009) Ensembl 2009. *Nucleic Acids Res.*, **37**, D690–D697.
5. Kersey,P.J., Duarte,J., Williams,A., Karavidopoulou,Y., Birney,E. and Apweiler,R. (2004) The International Protein Index: An integrated resource for proteomics experiments. *Proteomics*, **4**, 1985–1988.
6. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
7. Chen,C., Natale,D.A., Finn,R.D., Huang,H., Zhang,J., Wu,C.H. and Mazumder,R. (2001) Representative proteomes: a stable, scalable and unbiased proteome set for sequence analysis and functional annotation. *PLoS One*, **6**, e18910.
8. Jain,E., Bairoch,A., Duvaud,S., Phan,I., Redaschi,N., Suzek,B.E., Martin,M.J., McGarvey,P. and Gasteiger,E. (2009) Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics*, **10**, 136.

9. Goujon,M., McWilliam,H., Li,W., Valentin,F., Squizzato,S., Paern,J. and Lopez,R. (2010) A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res.*, **38**, W695–W699.
10. Patient,S., Wieser,D., Kleen,M., Kretschmann,E., Martin,M.J. and Apweiler,R. (2008) UniProtJAPI: a remote API for accessing UniProt data. *Bioinformatics*, **24**, 1321–1322.
11. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,U., Daugherty,L., Duquenne,L. *et al.* (2009) InterPro: the integrative protein signature database (2009). *Nucleic Acids Res.*, **37**, D224–D228.
12. Vizcaíno,J.A., Côté,R., Reisinger,F., Foster,J.M., Mueller,M., Rameseder,J., Hermjakob,H. and Martens,L. (2009) A guide to the proteomics identifications Database proteomics data repository. *Proteomics*, **9**, 4276–4283.